2019

Using Word Embeddings to Explore the Language of Depression on Twitter

Sandhya Gopchandani
University of Vermont

Using Word Embeddings to Explore the Language of Depression on Twitter

A Thesis Presented

by

Sandhya Gopchandani

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Master of Science
Specializing in Computer Science

May, 2019

Defense Date: March 22nd, 2019
Dissertation Examination Committee:

Christopher Danforth, Ph.D., Advisor
Kelly Rohan, Ph.D., Chairperson
Peter Sheridan Dodds, Ph.D.
Laurent Hebert-Dufresne, Ph.D.
Cynthia J. Forehand, Ph.D., Dean of Graduate College

Abstract

How do people discuss mental health on social media? Can we train a computer program to recognize differences between discussions of depression and other topics? Can an algorithm predict that someone is depressed from their tweets alone? In this project, we collect tweets referencing "depression" and "depressed" over a seven year period, and train word embeddings to characterize linguistic structures within the corpus. We find that neural word embeddings capture the contextual differences between "depressed" and "healthy" language. We also looked at how context around words may have changed over time to get deeper understanding of contextual shifts in the word usage. Finally, we trained a deep learning network on a much smaller collection of tweets authored by individuals formally diagnosed with depression. The best performing model for the prediction task is Convolutional LSTM (CNN-LSTM) model with a F-score of 69% on test data. The results suggest social media could serve as a valuable screening tool for mental health.

Dedication

To Mumi, Papa, Bha and Komi for always believing in me.

Acknowledgements

The outcomes of this project required a lot of guidance and assistance from many people and I am extremely grateful to have got this all along the completion of my thesis. I respect and thank my advisor, Chris Danforth for his support and guidance throughout this project. I would like to thank Andrew Reece for imagining this idea and trusting in me to do it, Ryan Gallagher and Peter Dodds for their feedback and suggestions whenever I needed it. In addition to that, thanks to Ben and Dave for helping me with understanding of concepts that was helpful in moving forward. Thank you to Rakesh, my brother for being readily available to answer my questions and confusions, to Viktoria for her warm hugs, to Kristin for checking on me and being a supportive friend. Finally, many thanks to Ahsan for providing constant reassurance and encouragement throughout this project.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Depression was considered a spiritual condition caused by demons and evil spirits until the seventeenth century [3]. The condition we now call "depression" was known as melancholia and was often treated with methods as beatings, physical restraint, and starvation in an attempt to let evil demons out [4]. It was due to a Greek physician named Hippocrates who associated depression with the imbalance in chemical fluids in human body called humours in his Aphorisms [5]. He used bloodletting, baths, exercise, and diet to treat depression. In 1621, Robert Burton outlined both social and psychological causes of depression in his book called Anatomy of Melancholy [6]. He considered fear, poverty and loneliness as causes of depression and recommended diet, exercise, travel and music therapy as a treatment for this condition.

Remedies to treat depression were no more adequate in the late 19th century, so people with severe depression were treated with lobotomy [7], a surgical procedure to destruct the front portion of brain, which seemed to provide calming effects. But such treatment was proven unsuccessful, causing personality changes, inability to make decisions, poor judgment, and sometimes death. It was only in the 1950s

when a tuberculosis medication called isoniazid was found to be helpful in treating depression in some people and that resulted in development of drug therapy as a possible choice to treat depression [8].

According to a World Health Organization (WHO) report, depression is the leading cause of disability, affecting about 300 million people globally [9]. The rate of major depressive episodes is the highest among the individuals between 18 and 25. Moreover, the total economic burden of depression in the USA is estimated to be $210.5 billion in 2010 [10]. About 50% of it is attributed to workplace costs and decreased productivity, 45% to medical expenses and 5% to costs related suicide. Although there are effective treatments available to cure depression, less than 50% of the affected receive such treatments and only one in five receives treatment consistent with current practice. The reasons include lack of knowledge, unavailability of trained health care providers and social stigma attached with mental illness that has made individuals reluctant to take necessary treatment. [11].

Social media is emerging as a promising tool for detecting depression and analyzing the content of tweets has been a popular method to understand human behavior and mental illness [12–14]. There has been significant research that suggests user activity on social media can be helpful in inferring the key indicators of depression raising the possibility that social media could be used as a potential screening tool to detect different conditions in mental illness [15, 16]. But detecting depression on social media through different techniques in Natural Language Processing (NLP) is a complex task mainly due to the complicated nature of mental disorders. Moreover, indication of depression is often subtle and not obvious to the reader. These muted indicators in language which are not obvious to the human reader may be captured

by neural network architecture. By using the Twitter public data, we hope to create a neural word embedding model that is sensitive to signals present in language around depression. Acknowledging the complex ethical issues associated with algorithmic inference of mental health [17], as well as significant privacy concerns, the results of this project may be helpful in predicting depression in individuals consenting to let an algorithm read their tweets.

Chapter 2

Literature Review

A rich body of literature has explored social media as a lens through which we can understand and detect an individual's risk for mental illness. In 2013, De Choudhury et al. used clinically validated depression measures as well as Twitter activity to find individuals who had been formally diagnosed with depression [15]. In their study, they handcrafted features and fed into a statistical model that could predict if an individual had depression with 70% accuracy. In 2017, Reece et al. suggested improvements in previous study by incorporating tweets posted prior to the date of subjects first depression diagnosis [12]. Moreover, De Choudhury et al leveraged Facebook status updates along with survey data to predict postpartum depression in new mothers [18]. The model used demographic information, Facebook activity and linguistic expression to predict PPD with 35% accuracy. The research in this field is not only limited to text and survey data.

In another study, Reece et al. utilized participants's pictures on Instagram to reveal predictive markers of depression [16]. The data was collected through responses to a standardized clinical depression survey on MTurk. All these studies rely on

crowd-sourcing tools to hire volunteers to get access to individual social media feed and the amount of data is limited by those that can complete the appropriate survey [19].

In contrast to prior mentioned method, Coppersmith et al. used self-reported disclosure on Twitter to classify individuals suffering with Post Traumatic Stress Disorder, contrasting their usage with those who do not self-report such diagnoses [13]. Extending the same idea, one study collected Twitter self- reported diagnosis data for ten different mental conditions and built a machine learning classifier to separate users with conditions from control users based on age and demographics [14]. Another study utilized public post and comment data in Reddit mental health support communities to study the transition from mental illness to suicidal ideation in individuals [20]. The study was also able to distinguish between individuals likely to undergo the transition with high accuracy as opposed to who do not.

All these studies focused on predicting mental conditions at the individual level based on how one engages and expresses oneself on social platforms. We studied depression from the perspective of the language associated with it. Prior studies have indicated that mental health conditions show implicit changes in the language of affected individual in the form of shift in word usage or in word frequency [21–24]. The elevated use of word "I" [24], verbs in past tense [22] and absolutist words [23] in the language of depressed individuals are some of the examples. In our research, we present a novel approach to study language of depression on Twitter using neural word embeddings.

This technique is based on the concept of distributed word representation [25] rather than local representation and uses multiple neurons to represent a single word

to capture the dependency of words in the language [25]. In turn, it enables word representation to learn general concepts of language. Moreover, such distributed representations are capable of automatically capturing the predictive features from text, freeing researchers and practitioners from manually crafting and encoding specific features [26]. In the recent research, this technique has been extended to infer user embeddings [26, 27] and has been utilized for sarcasm detection [27] , irony detection [28] and content recommendation [29].

Chapter 3

Language of Depression

An individual's usage of words can give important clues about the aspect of their social psychological world [30]. In particular, depression is found to have its own language that is distinguishable from common vocabulary [31]. Several studies in the literature have attempted to understand the effects of mental health on the individual's language. [32]. One study in psychology suggests that use of pronouns and other small words in our vocabulary reveals the most about our personality, social skills and intentions [33]. While these words are less than 500 words in English vocabulary but they account for more than 50% of the words we speak, hear and read hence play an important part in how we express ourselves.

Another study employed computerized text analysis Linguistic Inquiry And Word Count (LIWC) [34] to analyze the language differences in the essays written by depressed and non-depressed individuals and found that the depressed individuals used first person pronouns more than non-depressed individuals [24]. In 2001, Stirman and Pennebaker conducted a study to examine word usage in suicidal and non-suicidal poets. The study shows that while both groups did not differ in the use of negative

words, suicidal poets used more first person pronouns (I, me, myself) compared to collective words (we, us, ourselves) [35]. This might suggest that people with mental illness tend to focus more on themselves while being less connected with others.

In another study, researchers found that people suffering from depression, anxiety and suicidal ideation exhibit elevated use of absolutist words (e.g always, never, completely) in their language compared to the people in controlled group [23]. The study also pointed out that absolutist words may be better indicator of mental illness compared to the negative emotion words as indicated in the past studies [36]. The research to study the effect of mental illness on language has employed essays [24], personal diaries and online blogs [37] and social media forums [23] but none of the studies have utilized the information available on popular social media platforms like Facebook and Twitter.

In this section, we want to use Twitter data of depressed individuals and non-depressed individuals to study the language of depression. We want to replicate the prior studies that showed increased used of first person singular pronouns and elevated use of absolutist words and want to find out if the results would also hold for Twitter data. In addition, we are also interested in whether the depressed individuals in our study would exhibit more use of negative emotion words as compared to non-depressed individuals. The motivation behind this study is to see whether the established linguistic patterns would hold for a microblogging [38] service like Twitter that only allows users to send messages of up to 140 characters.

3.1 Data

The data for this experiment was collected to conduct a prior research study about forecasting the onset of mental illness [12]. Participants in this data were recruited using Amazon's Mechanical Turk (MTurk) crowdsource platform. The dataset is comprised of 178207 tweets from 139 depressed individuals and 192809 tweets from 162 non-depressed individuals. Moreover, average number of tweets per individual in depressed group is 1282 while average number of tweets per individual in control group is 1190 as shown in table 3.1 below.

Table 3.1: Summary Statistics

	# tweets	# users	# $tweets_{median}$	#$tweets_{mean}$
Depressed	178207	139	853	1282
Controlled	192809	162	719.5	1190

Table 3.2: Summary Statistics: Threshold>=500 tweets per user

	# tweets	# users	# $tweets_{median}$	#$tweets_{mean}$
Depressed	168882	85	2354	1986
Controlled	181591	91	2306	1995

To do the language analysis on this data, we only considered individuals with at least 500 tweets. We did this to have enough content about each participant in our study. Table 3.2 shows the summary statistics after excluding users having less than

500 tweets. Our final dataset contains 168882 tweets from 85 depressed individuals and 181591 tweets from 91 non-depressed individuals.

3.2 Language Analysis

Figure 3.1 shows the frequency distribution of words in depressed and control group in log-log scale. The word with the highest frequency has a rank of 1 and the word with second highest frequency has a rank of 2 and so on. The amount of times a word appears is proportional to one over its rank. The distribution of words in two corpora of tweets seems to follow a heavy tailed distribution which is captured by Zip'f law [39] which suggests that the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word.

In order to compare the usage of pronouns, absolutist words and negative emotion words in both groups, we first combined all tweets per individual in both groups. Second, we calculated the frequency of each word appearing in an individual's vocabulary in each group. We did this to account for the fact that some individuals have more tweets than others and hence they have more presence on Twitter. Now, we have frequency of each word for each user in both the groups. In order to compare overall usage of words in two groups, we took an average over all the users in each group to find the mean frequency of each word in the group. We then used T-test to check whether the difference in the mean frequency of each word in two groups is statistically significant. We used p-value of 0.05 to check the significance which means that if p-value is less than or equal to 0.05, the difference between the mean

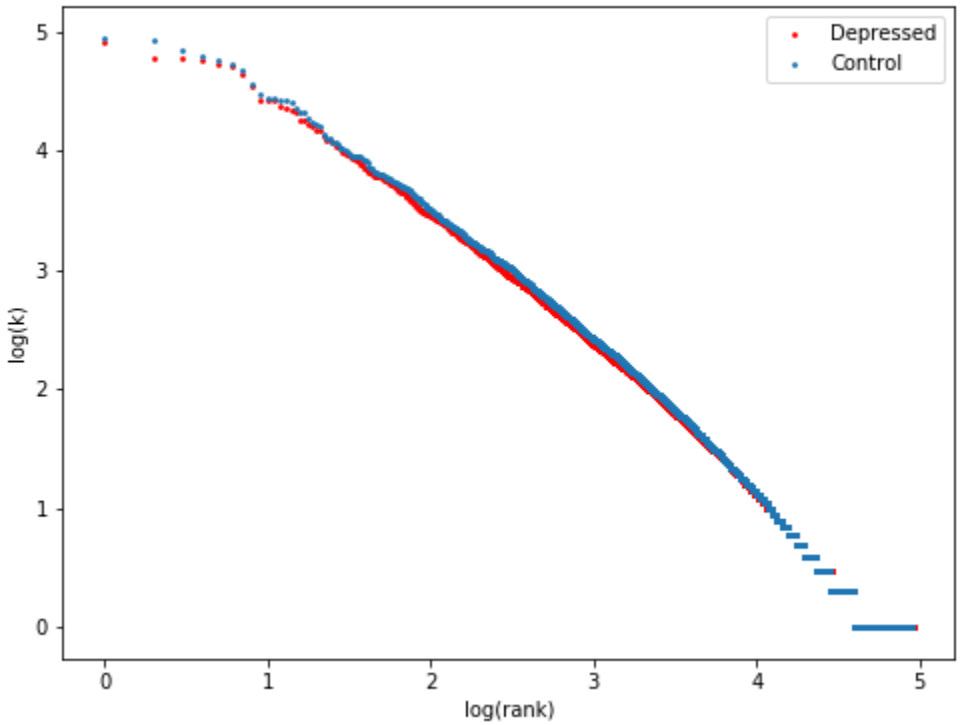

Figure 3.1: Rank versus Frequency for Words in Depressed Tweets and Non-Depressed Tweets

frequency of given word in each group is statistically different.

word	depressed(mean)	control(mean)	t-statistic	p-value	isFrequencyDifferent?
i	0.023799	0.023175	0.352221	0.725104	no
me	0.005552	0.004646	1.992282	0.047962	yes
myself	0.000371	0.000283	2.012806	0.045873	yes
mine	0.000175	0.000151	0.980858	0.328249	no
my	0.009758	0.008759	1.204894	0.229909	no
you	0.011580	0.011014	0.643397	0.520851	no
he	0.001509	0.001615	-0.477749	0.633512	no
she	0.000899	0.000866	0.269152	0.788147	no
it	0.007472	0.008093	-1.063378	0.289229	no
we	0.002172	0.002191	-0.076903	0.938797	no
they	0.001904	0.002038	-0.640315	0.522860	no
him	0.000748	0.000752	-0.031312	0.975059	no
her	0.001141	0.000993	1.106246	0.270170	no
us	0.000733	0.000648	0.919869	0.359117	no
them	0.001038	0.001012	0.254734	0.799236	no
yourself	0.000210	0.000162	1.872109	0.062966	no
himself	0.000058	0.000047	1.077993	0.282557	no
herself	0.000027	0.000027	-0.055725	0.955629	no
itself	0.000043	0.000034	0.948968	0.343987	no
ourselves	0.000022	0.000017	0.931266	0.353062	no
themselves	0.000065	0.000045	1.789204	0.075402	no

Figure 3.2: Frequency of Pronouns in Tweets from Depressed individuals and Non-Depressed individuals

Figure 3.2 shows the results for the list of pronouns. We can see that the mean frequency of first person pronouns(I, me, myself) in the depressed group is higher than the mean frequency of these words in the controlled group. But our t-test shows that the frequency of words 'me' and 'myself' is only statistically different in two groups that somewhat agrees to the past research [24, 40] that is depressed individuals use first person pronouns more than the non-depressed individuals but it does not hold true for the word 'I'.

word	depressed(mean)	control(mean)	t-statistic	p-value	isFrequencyDifferent?
absolutely	0.000110	0.000103	0.332394	0.739992	no
all	0.002973	0.002734	1.436829	0.152562	no
always	0.000683	0.000596	1.442692	0.151011	no
never	0.001059	0.000840	2.594357	0.010352	yes
nothing	0.000365	0.000360	0.149352	0.881450	no
totally	0.000194	0.000225	-0.833723	0.405620	no
complete	0.000068	0.000095	-1.691536	0.092815	no
completely	0.000107	0.000075	2.035807	0.043557	yes
definitely	0.000200	0.000238	-1.104618	0.270981	no
forever	0.000164	0.000121	1.793059	0.074856	no
constantly	0.000042	0.000029	1.708858	0.089260	no
entire	0.000125	0.000099	1.392173	0.165695	no
ever	0.000889	0.000824	0.777583	0.437875	no
every	0.000712	0.000586	2.259834	0.025147	yes
everyone	0.000499	0.000468	0.541826	0.588668	no
everything	0.000417	0.000357	1.275891	0.203862	no

Figure 3.3: Frequency of Absolutist Words in Depressed Tweets and Non-Depressed tweets

Figure 3.3 shows the results for the list of absolutist words. We can see that the mean frequency of few absolutist words like "never", "completely" and "every" are statistically different for depressed and non-depressed individuals using pvalue(<=0.05) but most of the absolutist words do not have different frequency distributions in two languages.

word	depressed(mean)	control(mean)	t-statistic	p-value	isFrequencyDifferent?
lonely	0.000032	0.000029	0.374511	0.708482	no
loneliness	0.000001	0.000001	-0.244697	0.806991	no
sad	0.000250	0.000242	0.249083	0.803594	no
sadness	0.000015	0.000008	1.445788	0.150119	no
fear	0.000073	0.000067	0.306158	0.759898	no
miserable	0.000021	0.000015	1.160137	0.247662	no
alone	0.000137	0.000123	0.688183	0.492327	no
helpless	0.000003	0.000003	-0.185728	0.852878	no
tired	0.000153	0.000258	-1.675177	0.096793	no
upset	0.000064	0.000058	0.441584	0.659343	no
no	0.002389	0.002197	1.045570	0.297313	no
not	0.003677	0.003465	0.777752	0.437773	no
feel	0.001034	0.000787	1.313791	0.191839	no
care	0.000310	0.000241	1.825165	0.070126	no

Figure 3.4: Frequency of Negative Emotion Words in Depressed Tweets and Non-Depressed Tweets

Figure 3.4 shows the results for the list of negative emotion words. It is interesting

to see that the mean frequency of given negative words is statistically same in depressed and non-depressed individuals implying that there is no significant difference in the usage of negative emotion words in languages of both groups.

The results of this experiment show that the past research to understand the effects of depression on language may not hold true specifically for the language used on Twitter. We find these results interesting because they lead to questions like whether the medium of language matters. Twitter being a microblogging platform allows user to share their thoughts in short and summarized way. Users may have to pick and choose the right words to get their message across which may influence how language might be effected by mental illness.

Chapter 4

Word Representations

4.1 Word vectors

According to DOMO's Data Never Sleeps 5.0 report [41], 90 percent of the data in the world was generated over the last two years alone. 456,000 messages are posted to Twitter every minute. There are 510,000 comments posted and 293,000 statuses updated every minute on Facebook. Everyday, we open Google and search for an article by providing few keywords and get hundreds of results in less than a second. Nate Silver analyzes million of tweets to correctly predict election results. We type in a sentence in Google and get it translated in various languages. All these tasks - clustering, classification and translation - require text processing. But computers do not understand text. So, we need some kind of numeric representation of text that machines can understand and process.

One could say that mapping a word to an integer might be a solution. Each word token would get an arbitrary integer and the vocabulary could be expanded as new words appear in the corpus. But this solution is as good as comparing if two words are

identical. Two words with related meanings might be assigned distant integers and two adjacent words assignment might have nothing to do with each other. So, there is no relationship in the words and information cannot be easily shared across words with similar properties. Moreover, this representation of words as unique, discrete id results in data sparsity. To overcome this problem, we need a representation for words that capture their meanings, semantic relationships and the different types of contexts they are used in.

In order to preserve the information around a word token, we can derive a notion of vector representation rather than integer representation for words where each dimension of the vector can be used for different purposes. For example, one dimension in the vector can represent the root each word belongs to: the words eat, eaten, eats, ate would be assigned 1 in the particular dimension while all other words that do not belong to this root would be assigned 0. This type of representation is referred as distributed word representation. It is based on the famous idea: You shall know a word by the company it keeps. [42]

4.1.1 Count-based Representations

Count-based methods are primarily based on the technique called Latent Semantic Analysis (LSA) [43] that uses documents as features of words based on the hypothesis that similar words would occur in similar documents.

Count Vector

Count Vector is a simple method that learns unique words from all the documents and then represents each word by counting the number of times it appears in each

document. Each dimension in the word vector represents the number of times it appears in each document.

TF-IDF Vectorization

TF-IDF is a frequency based method that not only takes into account the occurrence of a word in a single document but in the whole corpus of documents. It can be a considered a weighing method to extract important words in the document. It weighs down the common words appearing in almost all documents while giving more weight to less common words [44]. Term Frequency (TF) is a frequency of each word appearing in a document. The importance of a word increases proportionally to the number of times it appears in individual document. Inverse Document Frequency (IDF) is a measure of how common the word is across all documents. If a word appears in almost all the documents then it might not be important as compared to words that appear in few documents.

$$W_{i,j} = tf_{i_j} * log(N/df_i)$$

$W_{i,j}$ = weight of word i appearing in document j

tf_{i_j} = frequency of word i appearing in document j

N =Number of documents in a corpus

df_i = Number of documents where word i occurs

Co-Occurrence Matrix with Fixed Window Size

The co-occurence matrix tries to capture the semantic relationship between words by taking into account the neighboring words in a specified window size. The matrix is computed by counting how two or more words occur together in a given corpus. The resultant matrix can be very sparse depending on the size of vocabulary. If the vocabulary comprises of N unique words, co-occurence matrix will be of dimensions N by N. The technique is able to capture the powerful semantic and syntactic relationships in the vocabulary but it is computationally expensive due to high dimensionality. Therefore, techniques like Principle Component Analysis (PCA) [45] and Singular Value Decomposition (SVD) [46] are used to decompose co-occurence matrices to reduced dimensions with the least data-loss possible.

4.1.2 Prediction-based Representations

Predictive word representations are based on the idea that a words vector should encode its meaning by means of the neighboring words. Such representations of words use shallow neural networks to learn the word vectors as parameters of the model. The idea of predictive based word representation also called word embeddings have been around for some time [25, 47, 48] but it was in 2013 when Mikolov, Chen, et al. came up with two neural network based architectures to create high-dimensional word representations capturing semantic relationships between words unaided by external annotations. [1]

Continuous Bag-of-words CBOW Model

The Continuous Bag of Words (CBOW) is one of two model architectures introduced by Mikolov, Chen, et al in 2013. CBOW model learns word representations by probabilistic feedforward neural network to predict target word given the context words minimizing the following loss function:

$$E = -log(p(\vec{w_t}|\vec{W_t})$$

where w_t is the target word and W_t represents the sequence of words in the context. So loss function is negative log likelihood of target word given the context words.

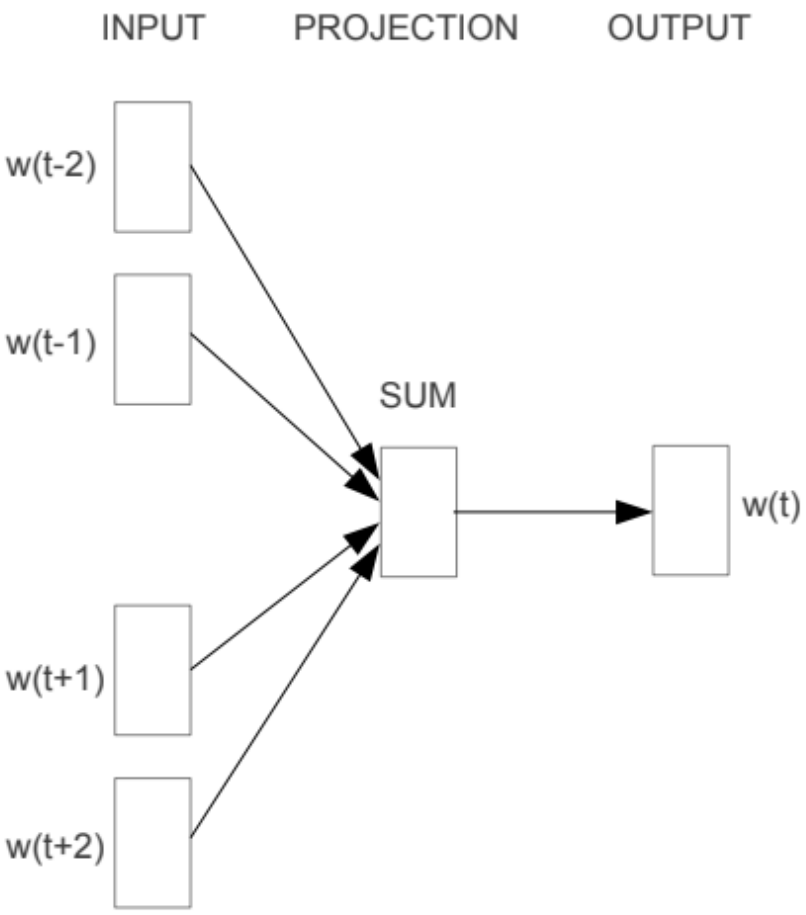

Figure 4.1: The CBOW Architecture Predicts the Current Word Based on the Context [1]

Skip-Gram Model

The skip-gram model is similar as CBOW but instead of predicting the target word based on context, it tries to predict the context words within a radius given the target word. So, skip-gram architecture reverses the use of target and center words. More details for the skip-gram model will be provided in a subsequent section.

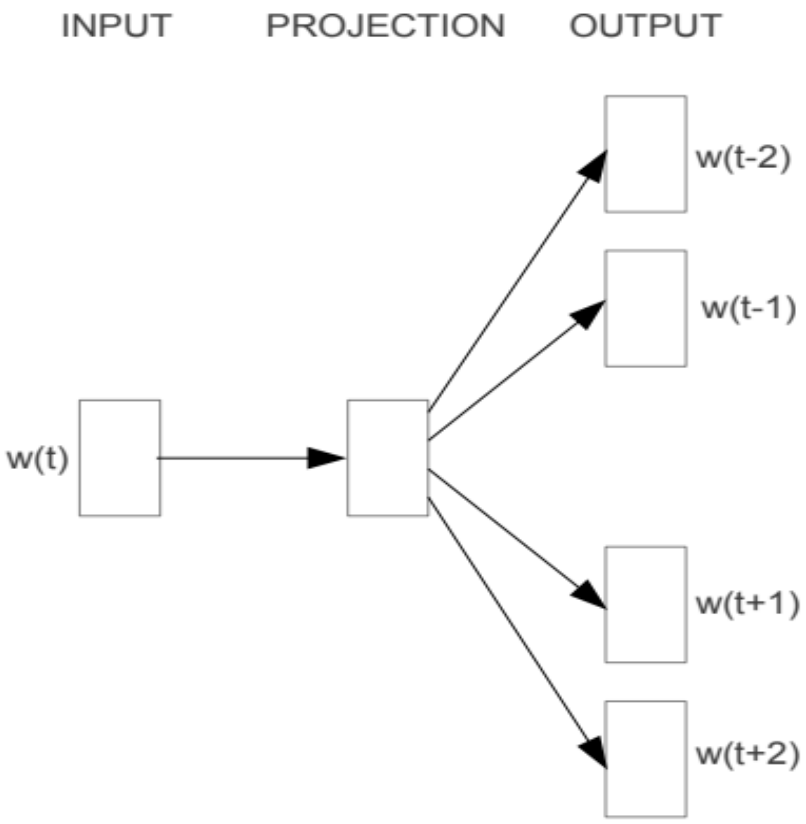

Figure 4.2: Skip-gram Architecture Predicts Context Words Given the Current Word [1]

4.2 Word2Vec with Skip-Gram Model

Word2vec is a neural network based framework that uses either CBOW or skip-gram model under the hood. The emphasis of word2vec model is to result in good word embeddings by using exceedingly large corpus of text data with lower computational complexity. The training of word2vec model does not involve dense matrix multiplication hence making the training very efficient [49]. In this study, we used the skip-gram model that has showed better performance in predicting infrequent words

given a large corpus [49]. The objective of the word2vec model is to have words with similar context occupy close spatial positions.

The idea of the Skip-gram model is to find low dimensional word representations that are good at predicting the neighboring words in the associated context. That is, at each step, we take one word as a target word and we try to predict the words in its context within a window radius of m. The model is going to define a probability distribution which describes the probability of the word occurring in the context given the target word. The objective of the Skip-gram model is to maximize the average log probability:

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{-m\leq j\leq m, j\neq 0} logp(w_{t+j}|w_t)$$

or minimize the negative log probability:

$$J(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{-m\leq j\leq m, j\neq 0} logp(w_{t+j}|w_t)$$

where T is number of words in the language (documents) and m is the size of training context. Larger m would imply more training examples hence more training time utilization. θ(s) are the parameters in the model that we want to optimize. And these θ(s) would be the vector representations for words. The idea is to keep adjusting θ parameters to minimize the loss function and in turn maximize the probability of the prediction. In skip-gram model, we define $p(w_{t+j}|w_t)$ using the softmax function as:

$$p(c|t) = \frac{exp(u_c^T v_t)}{\sum_{w=1}^{W} exp(u_w^T v_t)}$$

where v and u are "input" and "ouput" vector representations of w and W is the number

of unique words in the vocabulary. In the simple words, we take dot product of two word vectors that measures the similarity between two vectors. We use exponentiation to force the result to be positive and softmax function is employed to convert these numbers into probabilities. In practice, we use a different method to calculate $p(c|t)$ because cost of computing $p(c|t)$ is directly proportional to number of words W. Also Softmax is too expensive to use as a loss function as computing the gradients and optimizing millions of parameters (θs) in the model would make the training inefficient and slow.

In order to solve this issue, there are two optimization techniques discussed by the authors [49]. The technique which is the most popular and almost always used with skip-gram model is Negative Sampling. In the model, each training sample tweaks all of the parameters in the neural network which would make training very slow given the huge number of parameters. Negative Sampling solves this issue by having each training sample only update a small number k parameters in the model. The more detailed description of these methods can be found in [49].

So now that we have defined the loss function, we want to update the model parameters to minimize the loss function. And we do so by using Backpropagation using a Gradient Descent Algorithm that is used to optimize the parameters in the neural network by minimizing the loss. The standard way is to use chain rule to find the derivatives of loss function with respect to the input word vectors.

What is interesting about this method of embeddings compared to frequency based word embeddings is that this method does not assume each word as an independent token in the language rather it captures the context based on the neighboring words of the given word. The input to this model is a list of pair of words from tweets. Each

input is a word and output is the neighbor. Each pair is passed into a two layered neural network. Once the model is trained, we are interested in the learned parameters from the model which are word representations that supposedly have captured the context of the words in documents. Word vectors which are closer together should be more closely related than word vectors that are farther apart.

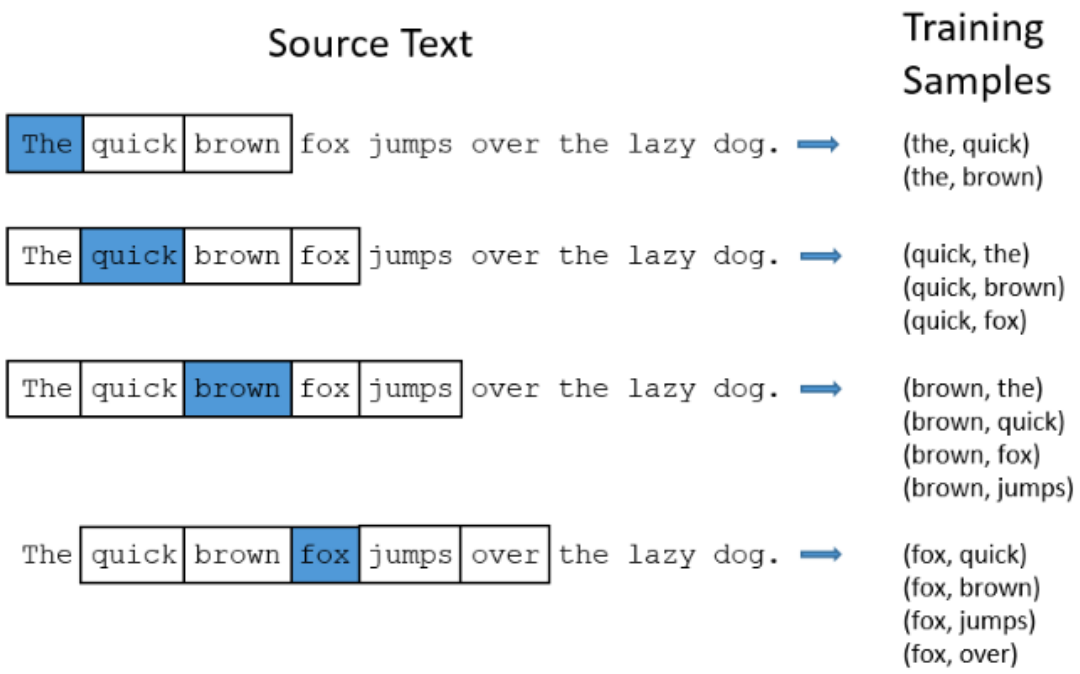

Figure 4.3: Word2Vec Skip-Gram - Training Example [2]

Figure 4.3 shows the structure of the training sample required for Word2Vec model. For each word in each tweet in the language, the algorithm takes pairs of the word and its neighbor. The number of neighbors is a parameter called window size that we specify when training the model. A window size of 2 would mean, we only want to consider two words on the each side of the input word. Usually, the smaller window size would capture the semantic similarity and large window size

would capture topical similarity among words. So, the context can be defined very locally or more broadly depending on the application of the model.

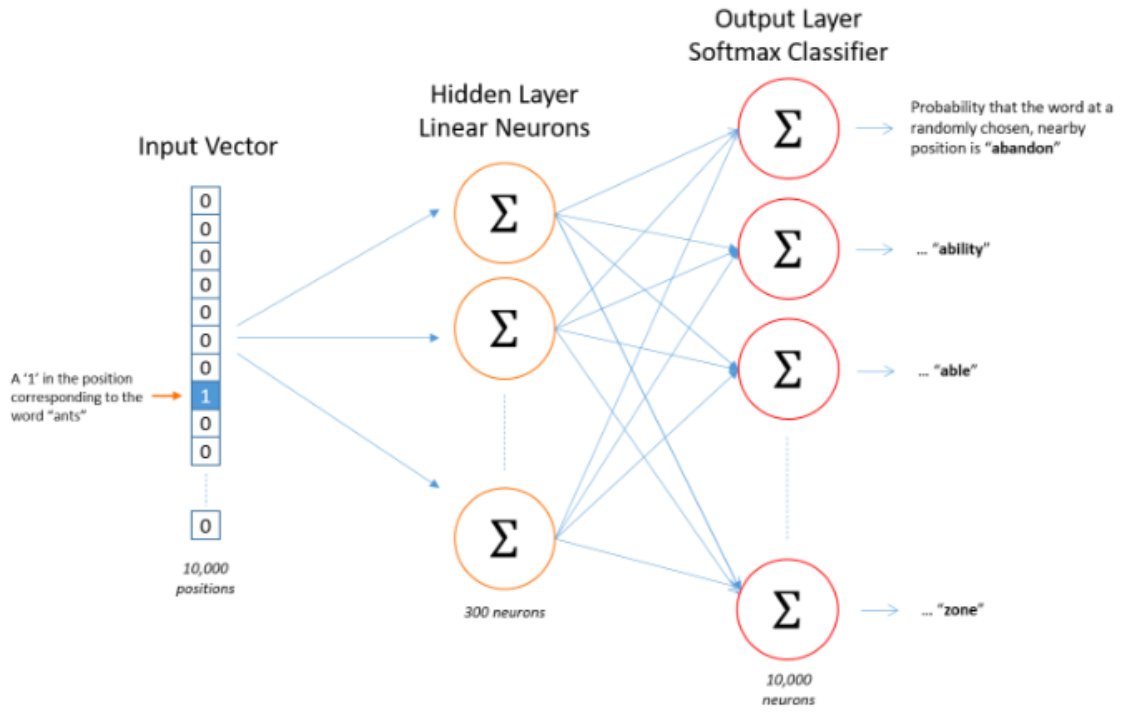

Figure 4.4: Skip-Gram Detailed Model Architecture [2]

Figure 4.4 shows the neural network architecture for Word2Vec with Skip-gram model. The input vector is of the size same as the total number of unique words in the vocabulary. The output of the network is a single vector, same size as input vector, containing the probabilities of word being the context word related to input word. There is no activation function in the hidden layer to introduce non linearity in the model as traditionally done in the neural network architecture. In this example, the hidden layer is represented by a weight matrix with 10,000 rows (one for every word in our vocabulary) and 300 columns (one for every hidden neuron). The rows of this weight matrix are actually the word vectors on interest. So, the end goal of this

network is to learn the weight matrix in the hidden layer by means of maximizing the probability of context word given the input word.

Chapter 5

Word Embeddings for Depression

So far, we have established that Word2Vec is a simple and scalable model which results in powerful word embeddings that can capture the semantic relationships in the language. In this study, we exploited this useful feature of word embeddings to study the language of depression and what kind of relationship the words would have in this language. The objective of this study was to train word embeddings to capture the differences between depressed language and general language. We defined "Depressed Language" as the set of tweets over the seven year period 2012-2019 that contains the words 'depressed' and 'depression' in them. We defined "General Language" as a set of random tweets. After collecting and preprocessing the dataset, we trained neural word embeddings using word2vec - one for each language corpus. The purpose was to train word vectors in each corpus and identify associations and equations that can be indicator of the hypothesis that use of language and context around the given word is different between two languages and hence can be a helpful tool to differentiate depressed language from healthy language.

5.1 Data

Word2vec is a data-hungry model that requires a large amount of data to learn the nuances in the language. For this purpose, we employed Twitter as our source of data. As of 2018, Twitter has 381 million active monthly users sharing 500 million tweets a day [50]. We used Twitter stream API Gardenhose that provides access to 10% of complete public tweets in near real time and has been used for many applications in the research studies [51–53]. Our dataset compromises of a subsample of 10% of tweets that mention the word "depressed" and "depression" in them. It contains 5.6 million tweets from 2012-2019 that define “Depressed Language” and 8.3 million random tweets that define “Healthy Language”. There is a difference in date range for two language groups to match the count of tweets in each corpus and also to limit computational time to train models. The imbalance between the two language corpora is obvious because there is a small percentage of tweets that talk about depression. The dataset is comprised of Json tweet objects. Each object contains several attributes related to a single tweet but we are only interested in text of the tweet for this study. It is a computationally challenging task to collect and work with the dataset of this size. So, we used multi-thousand-core, high-performance computing cluster Bluemoon of the Vermont Advanced Computing Core (VACC) to process and train word embeddings for two languages.

Table 5.1: Stats for Word2Vec Skip-Gram model

Dataset	# of Tweets	Vocabulary
Depressed	5.6 million	125.3K
Controlled	8.3 million	488.2K

5.1.1 Data Pre-processing

One of the challenges to work with the data extracted from Twitter is the unstructured nature of the text. Tweets posted by users contain misspelled words, new terms and syntax errors. So, data cleaning becomes an important step before training a model. Moreover, each tweet may contain links, abbreviation, keywords, emoticons and punctuation. So, it is crucial to streamline and clean the text in each tweet. In traditional Natural Language Processing (NLP) tasks, text is pre-processed by first stemming it, removing stop words, URLs, punctuation, numbers, emoticons and non-english words but this can result in potential loss of information and of language nuances that these tokens provide. So, for preprocessing, we adopted preprocessing script implemented at Stanford University [54]. In order to preserve the context and meaning of the text, words in the tweets are replaced with particular symbols to maintain the overall meaning of the tweet. Moreover, we do not remove stop words and pronouns because it has been shown that these words contain signals that capture the nuances in the language. Some of the examples of pre-processing is shown in the table below:

Table 5.2: Table 4.1 shows how each tweet is pre-processed and tokenized before training the model

Text	Symbol
$http://www.twitter.com/share$	$\langle URL \rangle$
$@anna$	$\langle USER \rangle$
2019	$\langle YEAR \rangle$
:)	$\langle SMILE \rangle$
: (	$\langle SADFACE \rangle$
#lifeisgood	$\langle HASHTAG \rangle lifeisgood$

5.2 Word Associations

Once we pre-processed the data, we trained two separate word2vec models on two languages - depressed and controlled - and got word embeddings for each word in two groups. Now that we have word representations, we can perform operations like finding similarity between two vectors. When words are represented as vectors, the similarity between two word vectors corresponds to the correlation between them. Cosine similarity is one of the most popular similarity measure applied to text documents [55] that provides an effective method to capture the linguistic and semantic similarity between the words vectors. The semantic similarity (S) of two word vectors (v1, v2) is then:

$$S(v1, v2) = cos(v1, v2) = \frac{v1.v2}{||v1||.||v2||}$$

And so the semantic distance between two word vectors can be calculated as:

$$D(v1, v2) = 1 - S(v1, v2)$$

In order to capture the differences between the two languages, we want to look at the associations of single word in the word embedding of two languages. By associations here, we mean which other words are similar to the input word in the vector space. The idea is that the words who appear together in the language tend to appear closer in the vector space and the cosine similarity between these word vectors should be closer to 1.

Word: workout

Depressed Language	General Language
workouts	workouts
gym	fitness
weightloss	exercise
practice	cardio
jogging	hiit
fitbit	yoga
fitness	muscle
eat	gym
squats	exercises
diet	weights
exercise	cooking
work	sesh
getfit	meal
fatloss	prep
skincare	bodybuilding

Figure 5.1: Top 15 Similar Words for the Word 'workout' in Two Languages

Figure 5.1 shows the lists of top 15 words similar to the word "workout" in two languages. It shows what other words people use or ideas they associate when they talk about "workout". While there are some associated words common in two languages but there are few words that stand out and are only common to depressed language. Words like workouts, fitness and gym have higher association with "workout" in both languages but words like "weightloss", "fatloss" and "skincare" are associated with "workout" only in depressed language suggesting that depressed language associates workout with different goal in the mind than people in general language.

Word: books

Depressed Language	General Language
novels	book
fanfics	authors
book	novels
essays	poems
chapters	writing
fics	reviews
manga	readers
magazines	poetry
fanfictions	writers
mangas	novel
fanfiction	fiction
imagines	literature
comics	letters
poems	reading
textbooks	ebook

Figure 5.2: Top 15 Similar Words For the Word 'books' in Two Languages

Figure 5.2 shows the top 15 word associations for the word 'books' in two languages. It is interesting to see how the words like fanfics, fanfiction and fanfictions all have high similarity score with respect to the word "books" in depressed language while these words do not appear in general language. Moreover, words like manga and comics are highly related to the word "books" only in depressed language and words literature, poetry and letters only appear in general language. In case it is not clear, manga are japanese comics and graphic novels created in Japan. The differences in how two languages associate the word "books" to separate ideas is what is captured

by these word embeddings which are able to capture the semantic of words in two languages.

Word: meals

Depressed Language	General Language
meal	snacks
sandwiches	meal
ramen	recipes
noodles	cocktails
snack	ingredients
snacks	lunch
nachos	lunches
pancakes	drinks
packets	specials
cupcakes	foods
burritos	snack
waffles	dishes
biscuits	salads
potatoes	seafood
tacos	eggs

Figure 5.3: Top 15 Similar Words for the Word 'meals' in Two Languages

Figure 5.3 lists the top 15 similar words for the word 'meals' in two languages. The word 'meal' is a general word and one would expect to see the similar context around this word in both languages but that is not the case as shown in the figure. Words like "ramen", "pancakes", "nachos", "potatoes", "tacos", "burritos", "biscuits" are most similar to the word "meals" in depressed language while none of these words appear in general language. Words like "cocktails", "eggs", "salads", "seafood", "lunch", "recipes" appear to be most related in the general language while they are missing in the depressed language. Our results show that depressed language

associates food with fast food which is high in fat and is related to obesity as shown in this study [56]. Past research shows that obesity is associated with depression [57]. Such differences point us towards fundamental differences in food consumption in two groups as established by previous research.

5.3 Yearly Contextual Shifts

This section is motivated by the idea that language evolves overtime. People and their ideas change as a consequence of social, political or personal change. In this section, we want to capture the contextual shifts in the usage of words in depressed language over the period of seven years (2012-2018). We want to examine how words might have developed varying linguistic context over the years. This would help us capture a yearly trend in the language that might help us understand depression better. For this experiment, we trained word2vec model for each year from 2012 to 2018. This means that each model will have subset of overall depressed language tweets. This may in turn affect the overall quality of word embeddings because of less data in each model. We excluded 2019 because we had only collected two month data for 2019 which we thought will not be representative of overall year. In order to capture the temporal changes in semantic shifts, we employed cosine similarity score of the words over the years. But given the stochastic nature of how word embeddings are trained, it might not be good idea to directly compare cosine similarity across models trained over different time period [58]. One way to align models and make them comparable is to allow for incremental updates of the model with new data without any modification. That is model trained on year y_{i+1} will be initialized with

the word embeddings from model from previous year y_i. This way, all the models are inherently related to each other which makes it possible to directly compare cosine similarities between the same word in different time period. This idea of incremental updates for temporal analysis was introduced by Kim et al. in 2014 [59].

Table 5.3: Yearly Stats for Word2Vec Skip-Gram model

Dataset	# of Tweets	Vocabulary
2012	938K	37,622
2013	1132K	39,784
2014	889K	35,583
2015	705K	32,127
2016	608K	31,836
2017	652K	35,365
2018	672K	36,771
2019	27K	4463

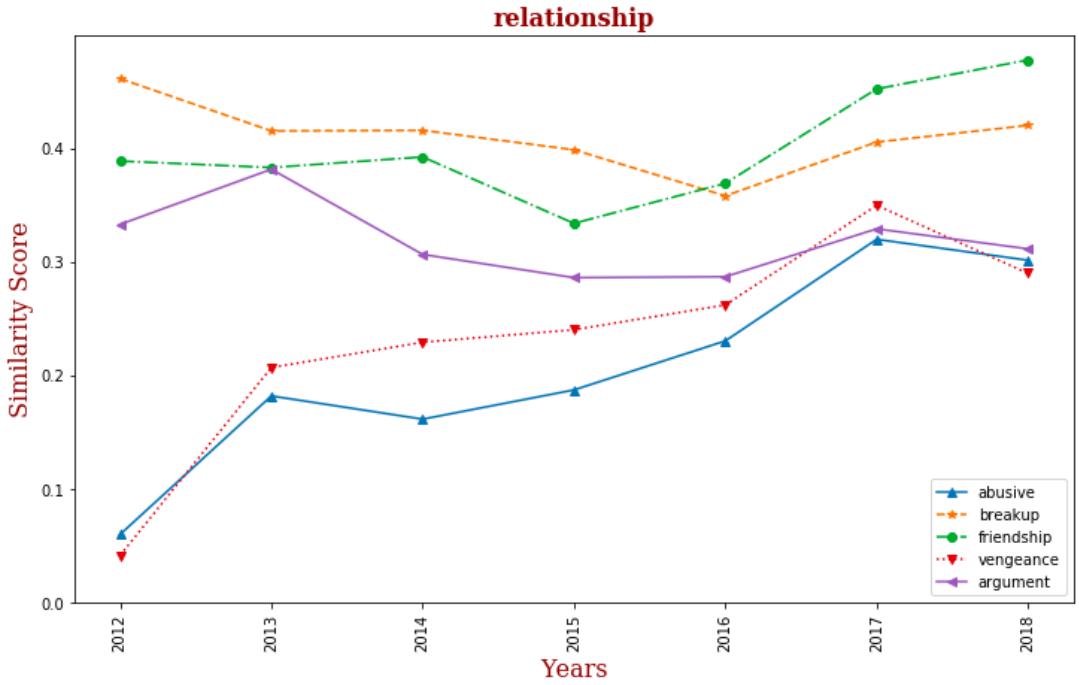

Figure 5.4: Contextual shift in the word 'relationship' over the past 7 years

Figure 5.4 shows the contextual shifts around the word "relationship" over the 7 year period. We define context as the subset of words having higher similarity score to the given word. We examine change in similarity score for context words over the seven year period which may give us some clue about how word is perceived now as compared to the past. It is worth noting how the word "abusive" and "vengeance" has increasing association with the word "relationship". It is also interesting to note that these context words follow the similar trend in each year for 7 years. This implies that that word "relationship" is being more often with the words "abusive" and "vengeance" over the time period.

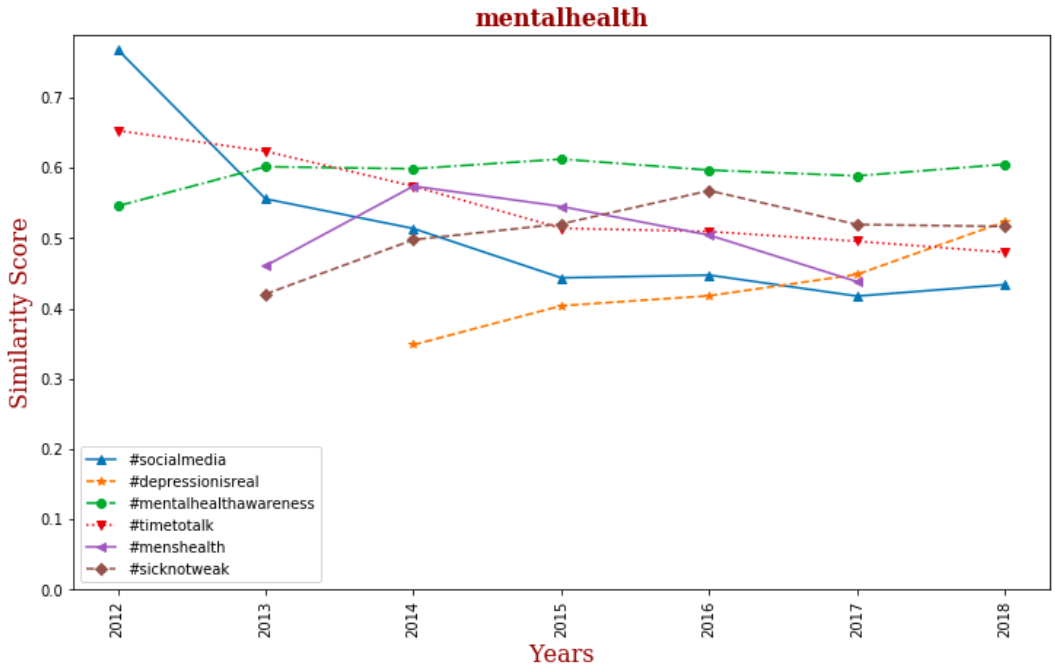

Figure 5.5: Semantic shift in the hashtag 'mentalhealth' over the past 7 years

Figure 5.5 shows the contextual change for a hashtag "mentalhealth". The figure shows some interesting results in the form of which hashtags around the topic of mental health have gained popularity over the years. The hashtag "Menshealth" (Men's Health) did not appear in the language when people talked about mental health in 2012 but it gained popularity in following years from 2013 to 2017. This means that men's health was not a topic of concern in earlier years rather it became a popular topic of mental health discussion in the following years. And that may have happened in the result of some social change taking place during those years. Moreover, the hashtag "sicknotweak" shows the similar trend. From non-existent in 2012, it has increasing similarity score in following years with respect to the given hashtag "mentalhealth". We can also find the similar trend for phrase "depressionisreal" that started to being part of mental health conversations in 2014 and continue to be so

with increased activity in the following years. It is worth noting that the hashtag 'socialmedia' shows decreasing trend over the time period.

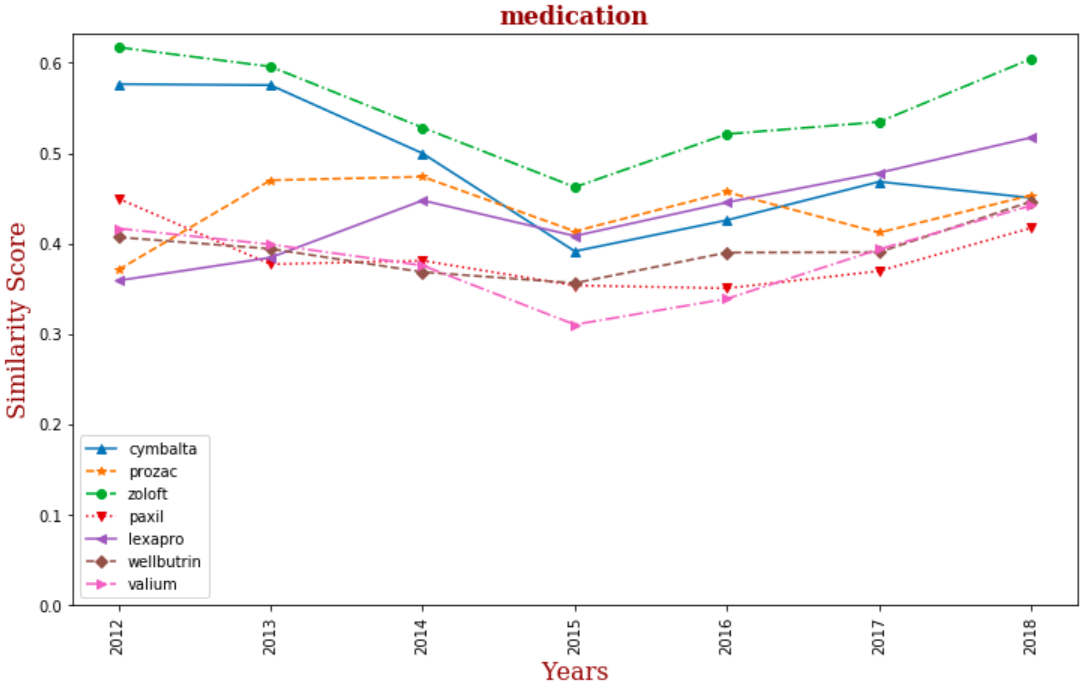

Figure 5.6: Semantic shift in the word 'medication' over the past 7 years

Figure 5.6 tries to capture the contextual change over the years for the word 'medication'. The idea was to capture if there is a trend in what medications are discussed more than the other in depressed language. We can see that cymbalta follows a decreasing similarity trend over the seven year period. The interesting thing about the results is that the medications zoloft, paxil, lexapro, prozac are antidepressant belonging to a group of drugs called selective serotonin reuptake inhibitor (SSRI) while cymbalta is also antidepressant but it belongs to a group called Serotonin norepinephrine reuptake inhibitor (SNRI). That might be indicator of whether there had been preferential shifts in anti-depressents usage over the years. One interesting thing

to notice here is that the jump in 2016 for 'lexapro' might be because of song Kanye West named Lexapro which was also released in 2016. Also, advertisement of these medications from bots on twitter might be responsible for the trend because we have not excluded the tweets from bots in our study.

Chapter 6

Prediction Task

Previous studies on predicting depression on social media have been done on an individual level [12, 15] where the task is to predict whether a person is depressed or not given her activity on social media. In this section, we want to predict depression at tweet level. In specific terms, we want to see whether a deep learning architecture with the use of pre-trained word embeddings is able to predict whether a tweet belongs to depressed individual or not based only on the text of the tweet. We employed various deep learning architectures namely CNN, LSTM and CNN-LSTM that have been widely used in classification tasks that are sequential in nature and involve natural language processing. We found that CNN-LSTM architecture achieve better results compared to CNN and LSTM model alone.

6.1 Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) were initially designed in the field of computer vision and have been widely applied to tasks involving visual data [60–62].

CNNs have the ability to recognize specific features inside a multi-dimensional field regardless of locality. The idea behind using CNNs on text data uses the fact that the language in the text is structurally organized and that we can expect a CNN network to learn important patterns in the text that would otherwise be lost. CNNs have been shown to determine discriminating phrases and hence are better suited to capture the semantics of the text [63].

6.2 Long Short Term Memory (LSTM)

Long Short Term Memory LSTM model is a type of Recurrent Neural Networks (RNNs) which are the state of the art algorithms for sequential data like time series, speech, text, financial data and weather. Such networks are designed to remember the previous state for given period of time [64]. LSTM model contains three gates: input, forget and output gate that control the flow of information from in and out of their memory. 'Input Gate' determine whether it should allow the information in the network, 'Forget Gate' determines how long certain information is in the memory and 'Output Gate' controls the impact of information in the memory to the output block of the network. Intuition behind LSTM for text analysis is that the network will remember the previous state in the text and thus will have a better understanding of the input.

6.3 CNN-LSTM Model

In this study, we combine the power of both architectures and propose a unified model CNN-LSTM for tweet classification. This architecture has shown to improve performance over individual CNN or LSTM models for text classification and sentiment Analysis [65, 66]. The intuition behind combining CNN and LSTM is that the convolution layer will extract local features and the LSTM layer will be able to use the sequence of those features to learn about the input. Figure 6.1 shows the architecture for CNN-LSTM model that we used for this study. The model takes word embeddings as input to the convolutional layer. The output of convolutional layer will be pooled into smaller dimension to reduce the parameter for training and also to extract relevant features. The output of that layer is fed into LSTM followed by output layer that predicts whether a tweet belongs to depressed individual or not.

6.3.1 Model Parameters

Parameters for our model were initially selected based on previous implementation of this model in text classification tasks. We further searched over parameter space through manual testing. Table 6.1 shows the parameters we selected for the final model.

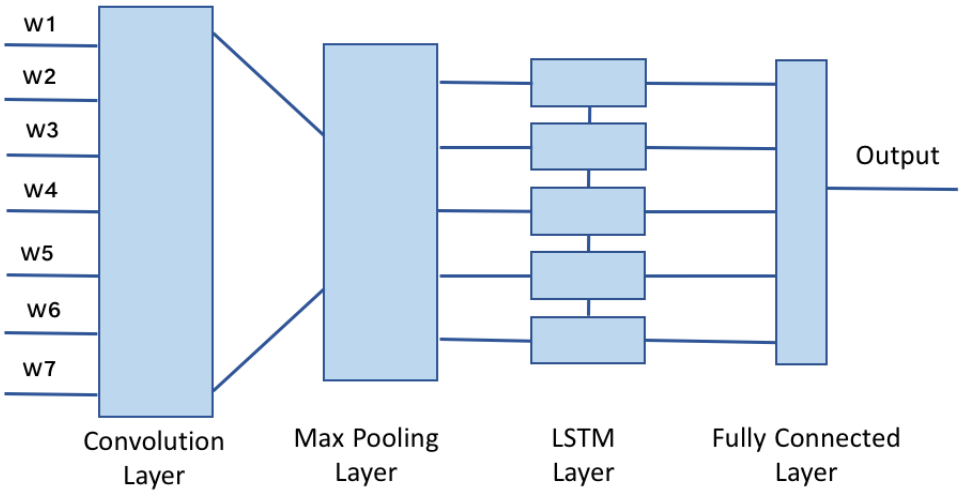

Figure 6.1: CNN LSTM Model Architecture

Table 6.1: Parameters for Model(s)

# Epoch	5
Batch Size	128
Filters	64
Kernel Size	3
Pool Size	2
Dropout	0.5

6.4 Results

In this section, we studied three neural network architectures CNN, LSTM and CNN-LSTM with three different variation for word embeddings. We trained these models by 1) initializing input word vectors by pre-trained word embeddings that we trained in chapter 5 and did not allow to update word embeddings during the training of the model(Pre-trained) 2) initializing input word vectors by random weights and allow for update through out the model training (Not Pre-Trained) 3) initializing input word vectors by pre-trained word embeddings but also allow for updates during the model training (Pre-trained + Trainable). We show the results of these models in Table 6.2.

Table 6.2: F1 Scores for Neural Network Models

# Embeddings	CNN	LSTM	CNN-LSTM
Pre-trained	0.38	0.46	0.56
Not Pre-trained	0.62	0.66	0.66
Pre-trained + Trainable	0.60	0.63	**0.69**

We can see that our CNN-LSTM model with pre-trained word embeddings along with update during the model training achieved an overall f1 score of 69% which is 3% higher than the same model with no pre-trained word embeddings and 13% higher than the same model with using only pre-trained word embeddings. The intuition behind this is the model has initial information about the input word data and then it goes on to update the parameters as it learns more about the language during the training. That implies that domain specific word embeddings would be helpful for

this prediction task.

This suggests that using pre-trained domain specific word embeddings might be a good way to boost the model accuracy. We can also see that using only pre-trained word embeddings in all three models results in much lower score. And that is because we do not allow the model to learn from input text data rather rely heavily on the word embeddings trained from a related yet different dataset which does not seem to be a good approach.

Table 6.3: Average Score of Different Neural Network Models

Neural Network Model	Average F1 Score
CNN	0.53
LSTM	0.58
CNN-LSTM	**0.64**

Table 6.3 compares three neural network models and reports an average F1 score for them. We can see that CNN-LSTM model achieved overall f1 score of 64% for this classification task which is 6% higher than regular LSTM model and 11% more than regular CNN model. This suggests that combination of CNN-LSTM architecture might be better suited for classification tasks involving NLP.

Chapter 7

Conclusion

In this study, we evaluated the semantic relationships captured by the words in depressed language and healthy language. We trained neural word embeddings using Word2Vec architecture to test whether the intuition captured by the model in one language differs from intuition in another language. We did so by comparing the relation terms in the nearest neighbors of a word in different word embeddings. We also looked at how context around words may have changed over the time to get deeper understanding of contextual shifts in the word usage. Finally, we trained a deep learning model to predict whether a tweet is from a depressed individual or a non-depressed individual. We found out that the domain of training corpus has a huge impact on the semantic relations represented by word embeddings. So, domain specific embeddings can be helpful in enhancing the performance of the complex tasks involving Natural Language Processing.

All NLP tasks require encoding of textual information to its numeric representation. Some word representations are based on the frequency distribution of word, based on the assumption that all words are independent in the vocabulary. More

recent methods are distributed vector representations based on the idea that the context of word can be understood by examining the company it keeps. Such representations mostly result from neural network based methods where the weights in the architecture are initialized randomly. We believe that the word embeddings trained on depression specific language can be used as an initialization step to represent word vectors in NLP tasks related to depression and perhaps mental illness. We think that these embeddings have captured the context that might be better representative of words in the corpus as compared to random initialization.

For this study, we assumed that depression related tweets and tweets from depressed individuals exhibit similar linguistic patterns. We did this because word embedding training requires a huge amount of data to correctly capture the semantic relationships in the vocabulary. And we do not have access to this amount of data from subjects going through depression. We also assumed that language on Twitter is representative of language people use in real life which may or may not be case. Moreover, we think that we can enhance these embeddings by adding more domain specific data about the topic from variety of online platforms that may be able to better capture the nuances in the language and may help us generalize the language. Finally, we believe that this method of understanding intuition in two languages can be applied to virtually any two different groups and can be helpful in understanding the fundamental differences in the lens people perceive the world.

Bibliography

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[2] Chris McCormick. Word2vec tutorial-the skip-gram model, 2016.

[3] LEWIS R WILLMUTH. Medical views of depression in the elderly: historical notes. *Journal of the American Geriatrics Society*, 27(11):495–499, 1979.

[4] M Matews. How did pre-twentieth century theories of the etiology of depression develop. *Psychiatry On-line*, 2004.

[5] William Henry Samuel Jones, Edward Theodore Withington, Paul Potter, et al. *Hippocrates*, volume 1. Loeb Classical Library, 1923.

[6] Robert Burton. *The anatomy of melancholy*. JW Moore, 1857.

[7] Kucharski Anastasia. History of frontal lobotomy in the united states, 1935-1955. *Neurosurgery*, 14(6):765–772, 1984.

[8] Mark Dombeck Rashmi Nemade, Natalie Staats. Historical Understandings Of Depression, 2001.

[9] World Health Organization. Depression Fact Sheet, 2018.

[10] Paul E Greenberg, Andree-Anne Fournier, Tammy Sisitsky, Crystal T Pike, and Ronald C Kessler. The economic burden of adults with major depressive disorder in the united states (2005 and 2010). *The Journal of clinical psychiatry*, 76(2):155–162, 2015.

[11] Lisa J Barney, Kathleen M Griffiths, Anthony F Jorm, and Helen Christensen. Stigma about depression and its impact on help-seeking intentions. *Australian & New Zealand Journal of Psychiatry*, 40(1):51–54, 2006.

[12] Andrew G Reece, Andrew J Reagan, Katharina LM Lix, Peter Sheridan Dodds, Christopher M Danforth, and Ellen J Langer. Forecasting the onset and course of mental illness with twitter data. *Scientific reports*, 7(1):13006, 2017.

[13] Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, 2014.

[14] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10, 2015.

[15] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. *ICWSM*, 13:1–10, 2013.

[16] Andrew G Reece and Christopher M Danforth. Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6(1):15, 2017.

[17] Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent Silenzio, and Munmun De Choudhury. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency (Atlanta GA*, 2019.

[18] Munmun De Choudhury, Scott Counts, Eric J Horvitz, and Aaron Hoff. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 626–638. ACM, 2014.

[19] GACCT Harman and Mark H Dredze. Measuring post traumatic stress disorder in twitter. *In ICWSM*, 2014.

[20] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110. ACM, 2016.

[21] Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *arXiv preprint arXiv:1804.07000*, 2018.

[22] D Smirnova, E Sloeva, N Kuvshinova, A Krasnov, D Romanov, and G Nosachev. 1419–language changes as an important psychopathological phenomenon of mild depression. *European Psychiatry*, 28:1, 2013.

[23] Mohammed Al-Mosaiwi and Tom Johnstone. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, page 2167702617747074, 2018.

[24] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133, 2004.

[25] GE Hinton, JL McClelland, and DE Rumelhart. Distributed representations, parallel distributed processing.–explorations in the microstructure of cognition, vol. 1. *Foundations*, 1986.

[26] Jiwei Li, Alan Ritter, and Dan Jurafsky. Learning multi-faceted representations of individuals from heterogeneous evidence using neural networks. *arXiv preprint arXiv:1510.05198*, 2015.

[27] Silvio Amir, Byron C Wallace, Hao Lyu, and Paula Carvalho Mário J Silva. Modelling context with user embeddings for sarcasm detection in social media. *arXiv preprint arXiv:1607.00976*, 2016.

[28] Cynthia Van Hee, Els Lefever, and Véronique Hoste. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, 2018.

[29] Yang Yu, Xiaojun Wan, and Xinjie Zhou. User embedding for scholarly microblog recommendation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 449–453, 2016.

[30] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

[31] Wilma Bucci and Norbert Freedman. The language of depression. *Bulletin of the Menninger Clinic*, 45(4):334, 1981.

[32] David E Losada and Fabio Crestani. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39. Springer, 2016.

[33] James W Pennebaker. The secret life of pronouns. *New Scientist*, 211(2828):42–45, 2011.

[34] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[35] Shannon Wiltsey Stirman and James W Pennebaker. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic medicine*, 63(4):517–522, 2001.

[36] Brendan Bradley and Andrew Mathews. Negative self-schemata in clinical depression. *British Journal of Clinical Psychology*, 22(3):173–181, 1983.

[37] Aubrey J Rodriguez, Shannon E Holleran, and Matthias R Mehl. Reading between the lines: The lay assessment of subclinical depression from written self-descriptions. *Journal of personality*, 78(2):575–598, 2010.

[38] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[39] Wentian Li. Random texts exhibit zipf's-law-like word frequency distribution. *IEEE Transactions on information theory*, 38(6):1842–1845, 1992.

[40] Nicholas S Holtzman et al. A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality*, 68:63–68, 2017.

[41] INC DOMO. Data never sleeps 5.0, 2017.

[42] John Rupert Firth. A synopsis of linguistic theory 1930-1955 in studies in linguistic analysis, philological society, 1957.

[43] Thomas K Landauer, Peter W Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

[44] Thomas Roelleke and Jun Wang. Tf-idf uncovered: a study of theories and probabilities. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 435–442. ACM, 2008.

[45] Lindsay I Smith. A tutorial on principal components analysis. Technical report, 2002.

[46] Virginia Klema and Alan Laub. The singular value decomposition: Its computation and some applications. *IEEE Transactions on automatic control*, 25(2):164–176, 1980.

[47] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[48] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[49] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[50] C Newberry. Twitter statistics all marketers need to know in 2018. *Hootsuite Blog, Retrieved October*, 14:2018, 28.

[51] Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media*, 2011.

[52] Mustafa Sofean and Matthew Smith. A real-time architecture for detection of diseases using social networks: design, implementation and evaluation. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, pages 309–310. ACM, 2012.

[53] Sounman Hong and Daniel Nadler. Which candidates do the public discuss online in an election campaign?: The use of social media by 2012 presidential candidates and its impact on candidate salience. *Government information quarterly*, 29(4):455–461, 2012.

[54] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[55] Ricardo Baeza-Yates, Berthier de Araújo Neto Ribeiro, et al. *Modern information retrieval*. New York: ACM Press; Harlow, England: Addison-Wesley,, 2011.

[56] ET Rolls. Understanding the mechanisms of food intake and obesity. *Obesity reviews*, 8:67–72, 2007.

[57] Chiadi U Onyike, Rosa M Crum, Hochang B Lee, Constantine G Lyketsos, and William W Eaton. Is obesity associated with major depression? results from the third national health and nutrition examination survey. *American journal of epidemiology*, 158(12):1139–1147, 2003.

[58] Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: a survey. *arXiv preprint arXiv:1806.03537*, 2018.

[59] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515*, 2014.

[60] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[61] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[62] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[63] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[64] Christopher Olah. Understanding lstm networks. 2015.

[65] Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. A c-lstm neural network for text classification. *arXiv preprint arXiv:1511.08630*, 2015.

[66] Mathieu Cliche. Bb_twtr at semeval-2017 task 4: twitter sentiment analysis with cnns and lstms. *arXiv preprint arXiv:1704.06125*, 2017.